DOCUMENT RESUME

ED 331 885 TM 016 445

AUTHOR Johnson, Eugene G.
TITLE Defining Levels on the 1990 Mathematics Composite.
PUB DATE Apr 91
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Cutting Scores; Equations (Mathematics); Grade 4; Grade 8; Grade 12; Intermediate Grades; Mathematical Models; Mathematics Achievement; *Mathematics Tests; Secondary Education; *Test Interpretation; *Weighted Scores
IDENTIFIERS Mapping; *Mathematics Composite (NAEP); *National Assessment of Educational Progress; Smoothing Methods

ABSTRACT

Procedures used to map the achievement levels, expressed as total expected scores on the full set of items presented to a grade, onto the Mathematics Composite of the 1990 National Assessment of Educational Progress (NAEP) are described. The Composite, defined as a weighted average of subdomain scores, provides a global measure of mathematics achievement. Levels of basic, advanced, and proficient were set for each of grades 4, 8, and 12 through a consensus process in which a panel of state and local educators, scholars, civic group representatives, and other citizens examined the questions and recommended the proportion that a student needed to answer to reach the defined achievement level. Expected test scores were mapped onto a unidimensional scale with five subscales for grades 8 and 12, and four subscales for grade 4. The process used to map achievement levels directly onto the mathematics composite included: (1) setting empirical item characteristic curves; (2) creating the empirical test characteristic curve; (3) smoothing the overall empirical test characteristic curve; and (4) mapping the cut points onto the composite. Results from the NAEP 1990 assessment had not been released at the time of this paper; thus, numerical results of the mapping are not provided. One table and two graphs illustrate the process. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED331885

DEFINING LEVELS ON THE 1990 MATHEMATICS COMPOSITE[1]

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Eugene G. Johnson
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
EUGENE G. JOHNSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

1. INTRODUCTION

In 1990 nearly 33,000 students participated in the National Assessment of Mathematics and an additional 110,000 eighth grade public school students from 40 states and jurisdictions participated in the Trial State Assessment of Mathematics. The goal of these assessments was to provide information to educational practitioners, policymakers, and the general public on the status of the mathematical performance of students. The basic information from the assessment is the responses of students to a large pool of items, selected via a consensus approach to measure how well students attain specified achievement goals during the course of their education. To provide concise and useful summarizations of this vast set of data, NAEP used item response theory based scales to capture the major findings of the assessment. To allow the measurement of student abilities within subdomains of mathematics, separate subscales were fit within each of a number of specified subdomains. To provide a global measure of mathematics achievement, NAEP has also produced a Mathematics Composite, defined to be a weighted average of the subdomain scores.

To have maximum utility, the points on the scale should be interpretable in terms of performance. For the 1990 Assessment of Mathematics, two distinct procedures were used to attach meaning to the overall Mathematics Composite. The first technique, scale anchoring, has been used by NAEP since 1984 and provides qualitative characterizations of performance at four successive points on the scale. The second technique, attempted for the first time in the 1990 Mathematics Assessment, sets specific achievement levels for each of the three grades, (4, 8, and 12), assessed. These achievement levels provide a statement of what students should know and be able to do at each of the three grades.

Three achievement levels, Basic, Advances, and Proficient, were set for each of the three grades 4, 8 and 12, in the assessment. Each of the achievement levels at each grade was decided via a consensus process in which a panel of state and local educators, scholars, employers, civic group representatives, and

---

[1]Paper presented as a part of the symposium: The NAEP Mathematics Scales -- A Useful And Innovative Way To Describe State And National Mathematics Achievement, American Educational Research Association 1991 Annual Meeting. The author is grateful to John Mazzeo and Donald Rock for enlightening discussions on the procedure to be used and to Alfred Rogers for the statistical programming.

TM016445

BEST COPY AVAILABLE

other interested citizens examined the assessment questions and recommended what proportion of questions students need to answer correctly to reach that achievement level. The end product of the achievement level setting process was a description of the types of tasks that students should be able to do to achieve a specific achievement level. A full description of the process used to define the achievement levels appears in Hambleton, Jones, and Bourque (1991).

To achieve maximum utility, the achievement levels had to be expressed as score points on the NAEP Mathematics Composite. This paper describes the procedures used to map the achievement levels, expressed as total expected scores on the full set of items presented to a grade, onto the Mathematics Composite. Unfortunately, since the results from the 1990 Mathematics Assessment had not been officially released at the time of the writing of this paper, the numerical results of the mapping of the achievement levels onto the mathematics composite cannot be given.

2. MAPPING ONTO A UNIDIMENSIONAL SCALE

Because the mathematics composite is an average of a series of subscales, standard IRT-based methods of mapping expected total test scores onto IRT scale scores do not apply. However, to set the scene and to define some concepts to be later generalized, the mapping of expected test scores onto a unidimensional scale will be considered first.

The scaling model used by NAEP for the mathematics assessment is the three-parameter logistic model from item response theory. This model quantifies students' tendencies to provide correct answers to the items going into the scale as a function of a parameter that is not directly observed, called proficiency and represented by the variable θ. The probability that a student whose proficiency on the scale is characterized by the unobservable variable θ will respond correctly to the item j is

$$P_j(\theta) = P(x_j=1|\theta, a_j, b_j, c_j) = c_j + (1-c_j)/\{1+\exp[-1.7a_j(\theta-b_j)]\} \quad (1)$$

where

- x_j is the response to item j, 1 = correct, 0 = incorrect
- a_j is the slope parameter of item j, characterizing the sensitivity of item to proficiency,
- b_j is the threshold parameter of item j, characterizing the difficulty of the item, and
- c_j for $0 \le c_j < 1$, reflects the chances of a correct response to the item by students of very low proficiency.

A plot of $P_j(\theta)$ as a function of θ is called the Item Characteristic Curve (ICC) for the item. The ICC for a given item gives both the probability of

success for a given value of θ and, more usefully in the current context, the value of θ such that the probability of success on the item is some specified value P.

In the case of a set of M items, all of which are assumed to fit a common unidimensional IRT model, equation (1) holds for all M items, where the item parameters are allowed to differ from item to item but where the proficiency θ is assumed to be common across items. Then, the common proficiency, θ, can be related to the expected score, T, across the M items via the Test Characteristic Curve (TCC):

$$TCC(\theta) = \sum_{j=1}^{M} P_j(\theta) \qquad (2)$$

Assuming that $TCC(\theta)$ is a monotonic function of θ, (as it will be when each of the $P_j(\theta)$ are monotonic), a unique value of θ can be found for any specified expected test score T.

3. PROPERTIES OF THE SUBSCALES AND THE MATHEMATICS COMPOSITE

Apart from issues relating to the assessment of the error in the prediction of θ for a given expected test score, equation (2) is sufficient for mapping expected test scores onto a unidimensional scale. However, because it was anticipated that different patterns of performance might exist within essential subdivisions of the subject area, the scaling for mathematics was carried out within the five mathematics contents areas specified by the Mathematics Objective Committee for the 1990 assessment. Five subscales were created for grades 8 and 12 and four subscales for grade 4. Four subscales common to all grades corresponded to the following content areas: Numbers and Operations; Measurement; Geometry; and Algebra and Functions. A Data Analysis, Statistics, and Probability subscale was also created for grades 8 and 12 but not for grade 4 because of an insufficient number of item in that content area for that grade.

In addition to the subscales, a mathematics composite was created. The composite allows for a global measure of achievement in mathematics while the subscale results provide diagnostic information about various facets of the subject area. The composite is defined as the weighted average of the subscale proficiency values, where the weights are proportional to the distribution of items by content area as specified by the Mathematics Objectives Committee.

Like all IRT scales, the subscales have a linear indeterminacy in that the mean and standard deviation of the subscale scores are arbitrary. The method that is used to resolve this indeterminacy has a direct impact on the value of the composite, and hence on the reported results. While a number of ways of setting the means and standard deviations of the subscales could be selected, we

chose to resolve the subscale indeterminacy by specifying that each subscale have a mean of 250.5 and a standard deviation of 50 over all three ages. For the Data Analysis, Statistics, and Probability subscale, which did not appear at grade 4, the means of the grade 8 and grade 12 students were matched to the average of the grade 8 and grade 12 means in the other four subscales. This produces a subscale proficiency score which can be interpreted in terms of the number of standard deviations that it is from the population mean (across the three grades). The composite score is similarly interpreted as a weighted average number of standard deviation units from the average population mean.

In NAEP, each individual responds to relatively few items within a given content area. Accordingly, NAEP does not compute IRT estimates of individual proficiency. Instead, NAEP constructs sets of plausible values designed to represent the distribution of proficiency in the population. A plausible value for an individual is not a scale score for that individual but may be regarded as a random draw from the distribution of potential scale scores for all students in the population with similar characteristics and identical patterns of item response as that individual. In NAEP analyses, five plausible values are associated with each student to allow evaluation of the effect of imprecision in individual proficiencies on population estimates.

The fact that NAEP provides plausible values instead of individual proficiencies and that the mathematics composite is an average across subscales eliminated the direct use of equation (2) to map achievement levels onto the composite. A possible generalization of (2) to this situation is to:

1) develop a separate test characteristic curve for each subscale, where the curve for subscale k is $TCC_k(\theta_k)$, which is of the form of equation (2), but which assumes a different proficiency for each subscale;

2) have the judges define separate cut points for each subscale, where the cut point at a specific level for subscale k is T_k, say;

3) find the value of θ_k corresponding to T_k for each subscale; and

4) producing an overall proficiency level $\theta_.$ as the weighted average of the subscale proficiency levels:

$$\theta_. = \sum_k w_k \theta_k$$

where the weights w_k are the weights used to generate the composite from the subscales.

However, the present item pool was not specifically designed to support three standard setting levels at each of three grade levels, let alone at the subscale level where there were felt to be insufficient items within most subscales to provide results of sufficient generalizability. (Table 1 gives the number of items by grade and content area.) Instead, it was decided to generalize equation (2) so that an overall expected test score could be mapped directly onto the composite.

4. MAPPING ACHIEVEMENT LEVELS DIRECTLY ONTO THE MATHEMATICS COMPOSITE

A three-step process was employed to map achievement levels onto the mathematics composite. In the first step, empirical item characteristic curves were created, by grade, for each item which was presented to that grade as a part of the BIB-spiral portion of the assessment. An empirical item characteristic curve is a generalization of the item characteristic curve of equation (1) and relates the observed probability of success on an item to the composite θ. In the second step, empirical test characteristic curves (generalizations of the TCC's of equation (2)) were formed for each of the mathematics content areas and in the third step, an overall empirical test characteristic curve was formed as a weighted average of the content area empirical test characteristic curves. This curve was then smoothed and used to map the achievement levels. Details of the procedure are given below.

It should be noted that the original intention was to base the achievement levels on a broader set of items. In addition to the seven blocks of items presented in the BIB-spiral portion of the main assessment, three additional blocks, involving estimations and higher order thinking skills which were presented in the National Assessment but not in the Trial State Assessment (because they required tape recorder pacing). It was hoped that these items could be included into the achievement level setting process. However, it was also necessary that the goal level results be applicable to the results from the Trial State Assessment. Thus, it was initially hoped that achievement levels, based on total expected scores across the 10 blocks, could be mapped onto the mathematics composite, which is based on the 7 BIB-blocks at each grade. However, the estimation and Higher Order Thinking Skills items were ultimately excluded from the process since they did not go into the subscales or the composite and because there was an inadequate link between performance on those items and composite plausible values. It should also be noted that nine Data Analysis & Statistics items were included in the process for Grade 4, even though a Grade 4 Data Analysis & Statistics subscale was not created, since there was

an adequate link between performance on those items and the mathematics composite.

4.1. Setting Empirical Item Characteristic Curves (EICC):

The EICC of a given item i at a given grade gives an empirical estimate of the probability of success on the item as a value of the composite θ and is defined pointwise for each integral value of the composite θ between 100 and 400 by

$$EICC_i(\theta) = C_i(\theta) / N_i(\theta) \qquad (3)$$

where

$N_i(\theta)$ is the sum of the weights of all students responding to the item whose value of the composite is in the interval $\theta \pm d$ where d was set equal to 5. This included students who omitted the item or who responded "I don't know" but excluded students who did not reach the item. Each student came into the sum five times-once for each of the five plausible values associated with that student.

$C_i(\theta)$ is the weighted sum of the scores on the item of all students presented the item whose value of the composite is in the interval $\theta \pm d$. Consistent with the procedures used in NAEP IRT scaling, omits and "I don't know" responses to multiple choice items received a fractionally correct value equal to 1 over the number of choices (excluding the "I don't know choice"). Omits for open ended items were counted as wrong (value 0). Each student came into the sum five times-once for each plausible value.

For overlap items presented to more than one grade, a separate empirical ICC was computed for each grade.

4.2. Creating the Empirical Test Characteristic Curve (ETCC):

Once the empirical ICCs were computed, an empirical test characteristic curve was computed as an average of the empirical ICCs of the items presented at a given grade. This was done by first computing empirical test characteristic curves for each of the five content areas and then determining the overall test characteristic curve for the grade as a weighted average of these.

For a given content area, the empirical content area test characteristic curve was defined pointwise for integer values of the composite by

$$ETCC_j(\theta) = \frac{1}{N_{\theta j}} \sum_{i=1}^{N_{\theta j}} EICC_i(\theta) \qquad (4)$$

where the sum runs over the $N_{\theta j}$ items classified as belonging to content area j for the grade which were reached by at least one student with a composite plausible value in the range $\theta \pm 5$.

The overall empirical test characteristic curve for the grade was a weighted average of the five content area empirical test characteristic curves:

$$ETCC_{overall}(\theta) = \sum_{j=1}^{5} w_j \cdot ETCC_j(\theta) \qquad (5)$$

where w_j is the relative weight for content area j determined by the learning area committee specifications and is given in Table 1.

Figure 1 shows the overall empirical test characteristic curve for grade 4. Because the results were embargoed at the time that this paper was written, the units on both axes have been suppressed. It can be seen that the curve is generally smooth except at the extremes.

4.3. Smoothing the Overall Empirical Test Characteristic Curve:

Because it is a statistic, the value of $ETCC_{overall}(\theta)$ for a given value of θ is subject to variability, the magnitude of which depends on the number of students with plausible values in the range $\theta \pm 5$. Because there are relatively fewer individuals with very high (or low) values of θ, the stability of the overall empirical curve deteriorates for θ in the upper (and lower) few percentiles of the proficiency distribution. The goal of smoothing was to reduce the variability of the test characteristic curve while producing a smoothed curve which was generally close to the empirical curve.

Figure 2 shows a plot of the logit of the ETCC as a function of θ. While the relationship between θ and the logit of the ETCC is nonlinear over the full range of θ, it was approximately linear for the upper range of values of θ. Accordingly, to handle the instability of the ETCC for large values of θ, while maintaining the empirical curve as much as possible, the following smoothing was conducted. First, a line was fit to the logits of the $ETCC_{overall}(\theta)$ for θ at and above the 95th percentile of the proficiency distribution for the grade, where the line was constrained to go through the logit of the ETCC at the 95th percentile. The resulting predicted values from this line were reconverted into

proportions and were used as the smoothed values of the ETCC for θ at or above the 95th percentile.

Then, to handle slight nonmonotonicities in the ETCC for smaller values of θ, the ETCC was subjected to a resistant smoothing algorithm designed to smooth away single points which are out of line with the surrounding values while preserving the overall pattern of the data. (The algorithm used is referred to as *4(3RSR)2H twice* in J.W. Tukey (1977).)

The final smoothed curve is denoted $ETCC_{overall}^{smoothed}(\theta)$ and is shown as the line on Figure 1.

4.4. Mapping the cut points onto the composite:

The final step involved finding the value of θ such that, given an achievement level cut point A (expressed as a percent correct across all items given to the grade), $ETCC_{overall}^{smoothed}(\theta) = A$. Linear interpolation was used for this mapping, with the final value of θ being rounded to whole units. Because the results are embargoed at this time, the achievement levels and the corresponding proportions of students in the assessment achieving each level cannot be given in this paper.

5. CONCLUSION

The procedure given in the last section is a generally applicable solution to the problem of relating performance across one or more sets of items to any function of one or more proficiencies. Since the relationships between overall probability of success across the items and the composite of the proficiencies are empirically determined, the empirical test characteristic curve in not constrained to assume any form. Thus, this technique is applicable for situations where the assumed scaling model does not entirely hold. All that is required is a sufficiently strong relationship between the quantities being related. Thus, even though the items from the Data Analysis, Statistics, and Probability content area were not scaled for grade 4, these items could be included into the empirical test characteristic curve shown in Figure 1 because there was a sufficiently strong link at the student level between performance on those items and proficiencies within the other content areas.

An issue meriting further consideration is the stability of any mapping based on an empirical test characteristic curve. As noted earlier, the ETCC is a statistic whose stability depends on, among other things, the number of individuals with plausible values in each interval and the number and characteristics of the items going into the overall test score. Some indication of the stability of the composite value θ_A corresponding to a particular expected

test score A can be obtained by jackknifing the empirical test characteristic curve as follows:

1) Recompute the ETCC 5 times, each time leaving out a different one of the five content areas and averaging over the remaining four content areas. The ETCC based on leaving out the k^{th} content area is

$$ETCC^{-k}_{overall}(\theta) = \sum_{j \neq k} w_j \cdot ETCC_j(\theta) \Big/ \sum_{j \neq k} w_j$$

where $ETCC_j(\theta)$ is the empirical content area test characteristic curve for content area j as given in equation (4).

2) Let θ_A^{-k} be the composite value corresponding to expected test score A and using $ETCC^{-k}_{overall}$ as the empirical test characteristic curve.

3) The variance of θ_A is approximately

$$Var(\theta_A) = \frac{4}{5}\sum_{k} (\theta_A^{-k} - \theta_A)^2$$

In addition to measuring the variance due to the various content areas, $Var(\theta_A)$ also approximately) includes the variance due to sampling items within content areas and the variance due to sampling students (See Wolter, 1985, Section 2.4.5). Values of the standard error of θ_A (the square root of $Var(\theta_A)$) for the grade 4 data and for various values of A are:

Expected test score A:	0.70	0.75	0.80	0.85	0.90
Standard Error(θ_A):	3.3	3.6	5.2	5.8	9.2

It can be seen that the standard error of θ_A is much larger for expected test scores of .90 than for scores of .70. This is reflecting the greater instability of the empirical test characteristic curve at higher levels of θ, where there are relatively fewer individuals.

REFERENCES

Hambleton, R. K., Jones, R., & Bourque, M. L. (1991). *Setting Achievement Levels for the 1990 National Assessment of Mathematics*. Paper presented at American Educational Research Association 1991 Annual Meeting, Chicago.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass: Addison-Wesley.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: John Wiley & Sons.

Table 1: Number of items and Objectives Committee weights by subscale and grade

Content Area	Grade 4		Grade 8		Grade 12	
	# items	weight	# items	weight	# items	weight
Numbers and Operations	52	.45	46	.30	38	.25
Measurement	20	.20	21	.15	23	.15
Geometry	14	.15	26	.20	25	.20
Data Analysis, Stat. & Prob.	9*	.10	19	.20	22	.15
Algebra and Functions	14	.10	25	.20	37	.25
Composite	109		137		145	

* Excluded from composite scale for grade 4 but included in achievement levels

Empirical Test Characteristic Curve
for Mathematics Composite
Grade 4

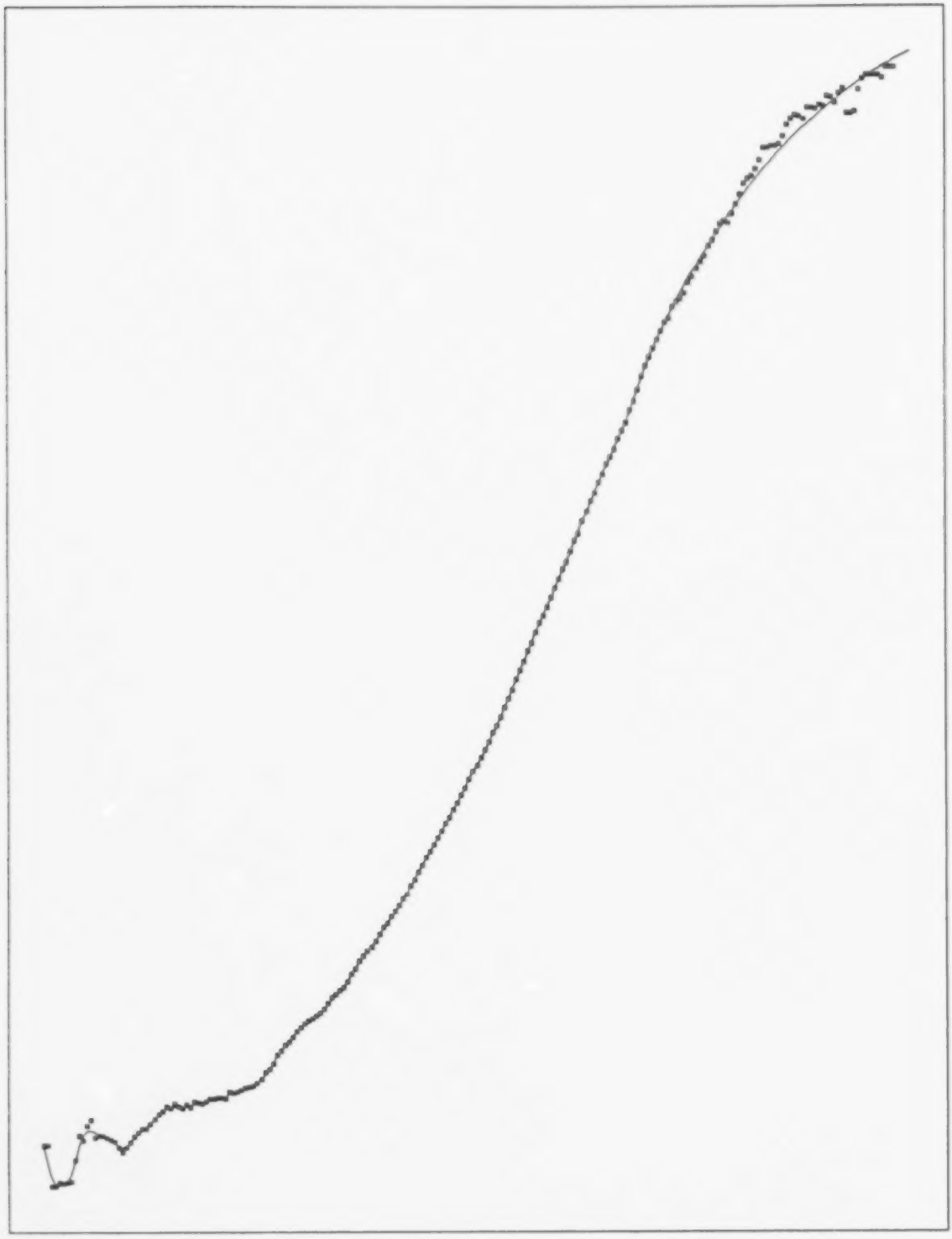

Figure 2

Logit of proportion correct vs theta
for Mathematics Composite
Grade 4